ACS Synthetic Biology

Open Access

pubs.acs.org/synthbio

Technical Note

Critical Assessment of Protein Engineering (CAPE): A Student Challenge on the Cloud

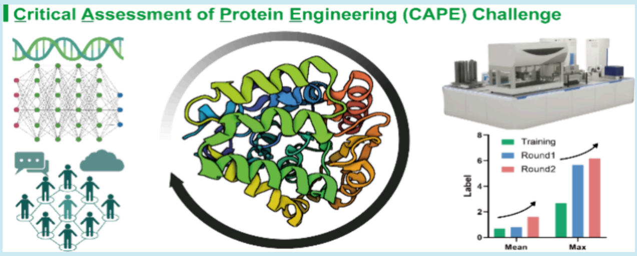
Lihao Fu, Yuan Gao, Yongcan Chen, Yanjing Wang, Xiaoting Fang, Shujun Tian, Hao Dong, Yijian Zhang, Zichuan Chen, Zechen Wang, Shantong Hu, Xiao Yi,* and Tong Si*

Cite This: https://doi.org/10.1021/acssynbio.4c00588

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The success of AlphaFold in protein structure prediction highlights the power of data-driven approaches in scientific research. However, developing machine learning models to design and engineer proteins with desirable functions is hampered by limited access to high-quality data sets and experimental feedback. The Critical Assessment of Protein Engineering (CAPE) challenge addresses these issues through a student-focused competition, utilizing cloud computing and biofoundries to lower barriers to entry. CAPE serves as an open platform for community learning, where mutant data sets and design algorithms from past contestants help improve overall performance in subsequent rounds. Through two competition rounds, student participants collectively designed >1500 new mutant sequences, with the best-performing variants exhibiting catalytic activity up to 5-fold higher than the wild-type parent. We envision CAPE as a collaborative platform to engage young researchers and promote computational protein engineering.



KEYWORDS: *protein engineering, student challenge, benchmark, machine learning, biofoundry*

INTRODUCTION

Understanding the sequence−structure−function relationship is crucial for studying and engineering proteins, the workhorses of life's myriad biological processes. The groundbreaking success of AlphaFold highlighted the impressive achievement of machine learning in sequence−structure prediction with high accuracy.[1] However, AlphaFold predictions are largely confined to static structures, whereas protein function requires complex, dynamic conformational changes. Therefore, computational sequence−function prediction remains an unmet challenge, limiting our capability to design and engineer improved proteins for diverse applications in medicine, agriculture, energy, and chemical production.

The major obstacle in protein function prediction and engineering by machine learning lies in the scarcity of sizable, high-quality data sets for model training.[2,3] For enzymes, in particular, the broad range of catalytic mechanisms, reaction types, experimental conditions, and reporting formats prevents robust pooling of function data, in stark contrast to the more standardized databases for protein sequences and structures. Moreover, a primary goal of function prediction is to guide the design of protein sequences with new or improved properties. However, it remains time- and cost-intensive to experimentally validate algorithmically proposed sequences, which impedes rapid, iterative model refinement.

To address these challenges, we introduce the Critical Assessment of Protein Engineering (CAPE) challenge, which is inspired by the Critical Assessment of Structure Prediction (CASP) competition. CASP, initiated in 1994,[4] has been instrumental in driving progress in protein structure prediction by providing a platform for rigorous, blind assessment of prediction methods. Through biennial rounds of prediction and experimental validation, CASP has fostered a collaborative community and ultimately resulted in breakthroughs like AlphaFold.[1] Following this successful model, CAPE is designed as a student-focused competition to develop and benchmark protein engineering models while providing valuable educational experiences.

Unlike typical data contests with one-shot model evaluation, CAPE encompasses complete cycles of model training, protein design, laboratory validation, and iteration (Figure 1). To lower entry barriers for students and researchers at all levels, model training was performed on the Kaggle data science platform, while experiments were conducted in an automated biofoundry, both accessible to participants at no cost. The use of cloud-based resources and robotic assays ensures rapid feedback on model performance, unbiased reproducible

Received: August 29, 2024
Revised: October 24, 2024
Accepted: October 29, 2024

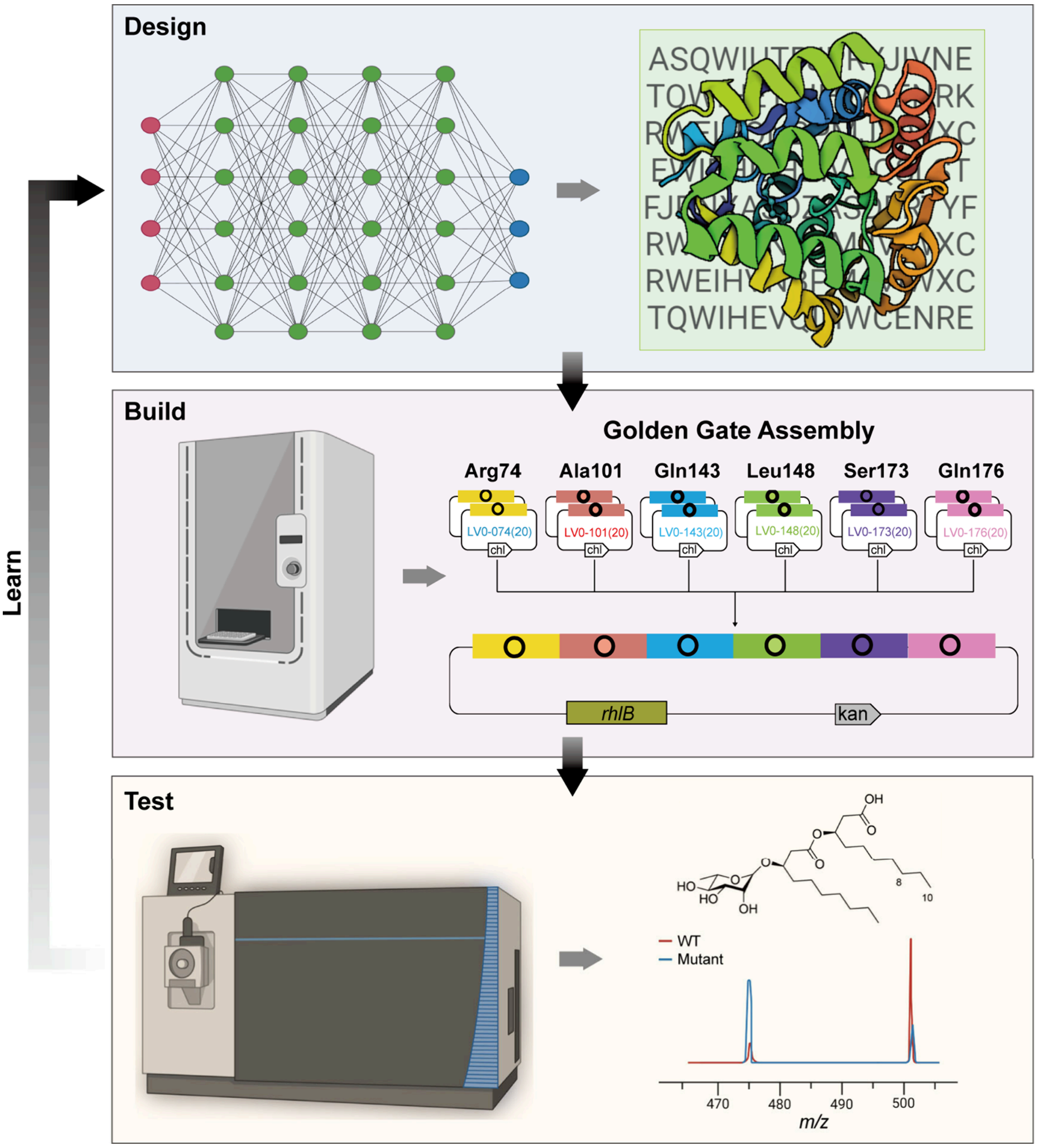


Figure 1. Design−build−test−learn (DBTL) cycle in the inaugural and second CAPE challenges.

benchmark, and equal opportunity for participants regardless of resource availability.[5−7] While previous student competitions have implemented some aspects of CAPE's vision, such as using cloud resources for community-driven innovation,[8−11] standardized assays for fair comparison across teams,[12] and biofoundry-assisted DNA synthesis,[7] CAPE has successfully integrated these desirable features, thanks to the recent advances in cloud computing and experimentation. This competition model can also be extended to other biological systems beyond proteins, as envisioned by the International Rational Genome-Design Contest (GenoCon) to engineer organisms with desired functions.[13,14]

In this technical note, we briefly summarize the design, execution, and results of the inaugural and second CAPE challenges, inviting community participation to solve protein engineering challenges using crowd-based creativity.

RESULTS AND DISCUSSION

Held from March to August in 2023 (Table S1), the inaugural CAPE Challenge focused on designing variant sequences of the RhlA protein, a key enzyme for producing rhamnolipids as an ecofriendly alternative to synthetic surfactants. This enzyme was chosen due to the availability of previously developed robotic protocols to create and screen mutant libraries on a biofoundry (Figure 1).[15] Participants were tasked with designing RhlA mutants with enhanced catalytic activity to increase rhamnolipid production in engineered *Escherichia coli* as a host. The challenge allowed modifications at up to 6 select positions in the RhlA amino acid sequence, with any of the 20 amino acids as options, presenting a design space of 20^6. Currently, we limit the design to combinatorial mutagenesis given the budget and turnabout constraints, where DNA assembly rather than DNA synthesis was utilized for

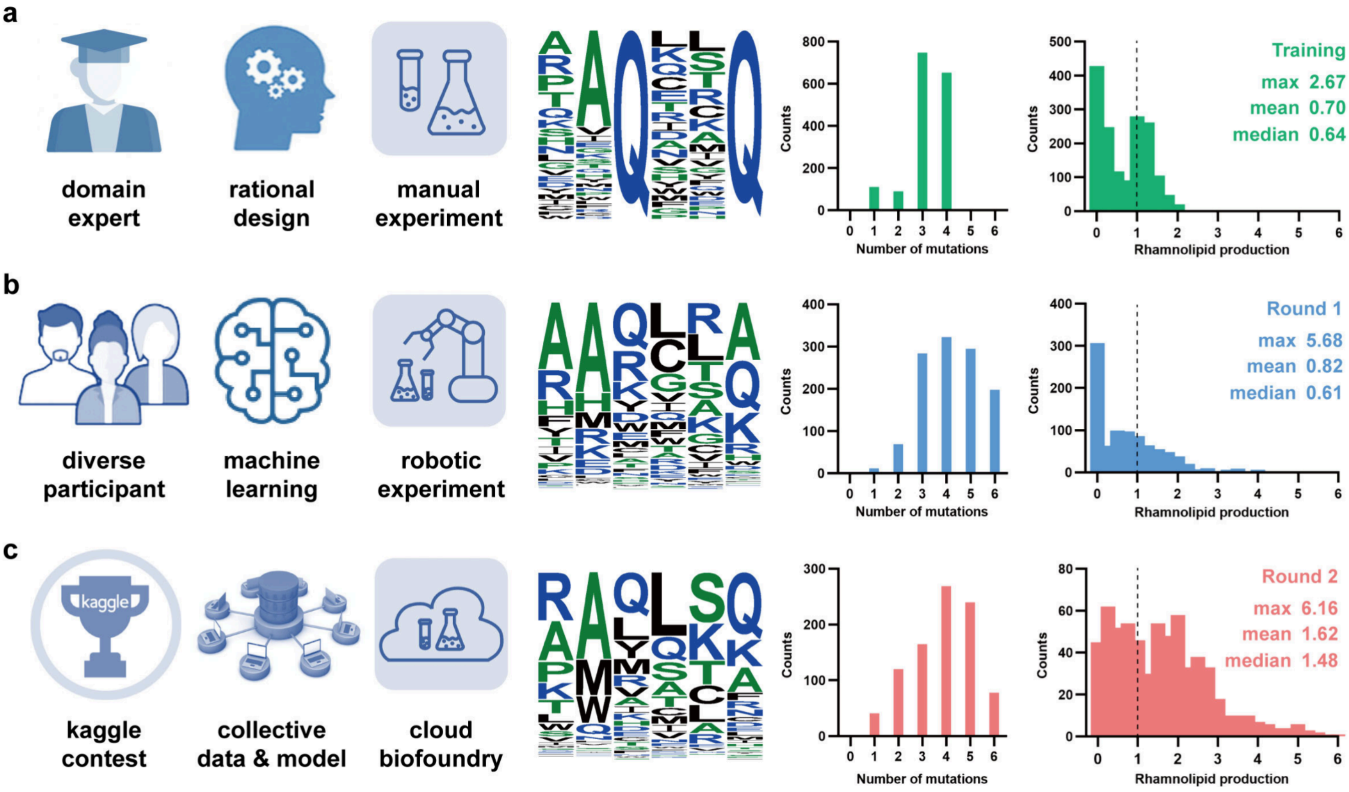


Figure 2. Data-centric overview of CAPE. Panels a−c correspond to the Training, Round 1, and Round 2 data sets, respectively, which were generated by an evolving combination of participants, algorithms, and experiments during the different phases of the CAPE challenge. The sequence logos and histograms summarize collective mutant sequence space, number of mutations per sequence, and function performance as rhamnolipid production (normalized to the WT level) of the corresponding data sets.

constructing any of the 64 million variant sequences within weeks (Figure 1).

Teams were given 1593 sequence−function data points for RhlA from previous research[15] (Figure 2a, the Training Set) to develop models, culminating in the submission of 96 variant sequences per team. To encourage new designs, submission of sequences that appeared in the training set was not allowed. After combining and removing replicates, 925 new enzyme sequences designed by 13 teams were physically built and tested (Figure 2b, the Round 1 set), using previously reported protocols on a biofoundry.[15] The scoring method reflects the goal of protein engineering to identify top-performing sequences, where the variants in the top 0.5%, 0.5−2%, and 2−10% ranges received 5, 1, and 0.1 points, respectively.

At the conclusion of the first CAPE challenge, a wealth of new sequence−function data was collected. Inspired by the CASP model of continuous model refinement and benchmarking, we wondered: could this newly generated data set serve as a resource for further model improvement in protein engineering? Rather than immediately releasing these data, we saw an opportunity to use it as a confidential test set for a second competition round.

The second CAPE challenge, held from September to December in 2023 (Table S1), was designed to contain two phases. The first phase was performed on the Kaggle data platform, where the original training set was provided to a new cohort of 25 teams, and the Round 1 results were used as a hidden evaluation set. Teams were allowed to repetitively refine models and submit predictions, which were automatically evaluated using Spearman's ρ on the Kaggle leaderboard.[16] Due to increased complexity and time commitment in sequence design, only 10 teams participated in the second phase, with each team submitting another 96 designs for validation. However, if in future competitions the number of sequence design submissions exceeds our experimental budget, we may limit participation in the second phase to the top-ranking teams from the Kaggle Leaderboard. In total, 648 unique, new sequences were collected and then experimentally assessed (Figure 2c, the Round 2 set). We applied the same scoring scheme to determine the winning team, eager to see whether this iterative approach would yield better designs relative to those in the first CAPE challenge.

To our excitement, student teams explored a diverse design space and engineered superior-performing RhlA mutants during the two successive CAPE challenges (Figures 2 and S1). The contestants achieved higher sequence diversities compared to the initial training set, as evidenced by the Shannon diversity index (Training, 2.63; Training + Round 1, 3.06; and Training + Round 1 + Round 2, 3.16). As the competition proceeded, a stepwise increase was achieved in the maximum, average, and median values of the protein functional performance. The best-performing mutants in the Training, Round 1, and Round 2 data sets produced rhamnolipid at levels of 2.67, 5.68, and 6.16 times that of WT production, respectively (Figure 2). Round 2 mutants showed greater improvements despite fewer proposed sequences in Round 2 (648) compared to Round 1 (925), indicating a higher success rate. Indeed, the iterative approach promoted collective learning, with the confidential test sets created collaboratively by competing teams and algorithms (Figure S1), reflecting the wisdom of the crowd.

The impressive performances of the Round 2 teams may be attributed to several factors. First, the data set expansion from 1593 to 2518 sequence−function pairs, coupled with increased sequence diversity (Shannon index rising from 2.63 to 3.06), likely enhanced model generalization and predictive accuracy. Second, unlike the training data set (Figure 2a), Round 1 contained higher-order mutants with five or six mutations (Figure 2b) that provided crucial information on nonadditive interactions among residues, enabling models to better capture complex epistatic effects. Moreover, the use of hidden test sets helped mitigate overfitting concerns.

Various predictive and design algorithms were devised by a diverse panel of participants. In the inaugural CAPE challenge (Table S2), the champion team was from Nanjing University and scored 29.1 points. They developed a deep learning pipeline that applied Weisfeiler−Lehman Kernel for sequence encoding, a pretrained language model for predictive scoring, and a coarse-grained scan combined with Generative Adversarial Network for sequence design (Figure S2a). For the second CAPE challenge (Table S3), the best-performing team on the Kaggle leaderboard was from the Beijing University of Chemical Technology, achieving a Spearman correlation score of 0.894 with an algorithm based on graph convolutional neural networks using protein 3D structures as the input (Figure S2b). However, this team's proposed sequences only ranked fifth in the experimental validation phase. In contrast, the Shandong University team won first place by applying a grid search to identify optimal multihead attention (MHA) architectures for positional encoding to enrich mutation representation (Figure S2c), although they ranked second in the Kaggle phase. The discrepancy in team ranking between the Kaggle and experimental phases is noteworthy and intriguing (Table S3). It underscores the persistent challenges in algorithm development for protein engineering, such as model overfitting and the difficulty of extrapolating beyond training data. Moreover, it highlights the crucial distinction between sequence-to-function prediction and the inverse problem of function-to-sequence design. This performance gap emphasizes that true algorithmic efficacy in protein engineering cannot be assessed by performance on existing data sets alone. The CAPE approach, with its integration of computational prediction and experimental validation, provides an essential benchmark for assessing the real-world applicability of these algorithms.

To facilitate retrospective analyses, all collected data sets and winning implementation are openly accessible on Github (https://github.com/KRATSZ/CAPE-2023). Our rudimentary analyses of the sequence encoding, predictive modeling, and mutant design strategies employed by CAPE participants (Tables S2 and S3) revealed several common approaches among the top-performing teams. These include ensemble methods combining multiple models, advanced encoding techniques incorporating structural and physicochemical information, attention-based architecture like transformers, and pretrained protein language models. While some teams performed full space enumeration, others opted for targeted exploration or stepwise filtering.

These winning approaches offer valuable insights for future CAPE participants and highlight key challenges in the field. The lack of high-quality, diverse data sets often leads to overfitting and poor generalization, while the vastness of sequence space makes comprehensive exploration computationally intractable. Furthermore, nonadditive interactions between mutations (epistasis) complicate model development, and models often struggle to extrapolate beyond their training data, a critical limitation when the goal is to identify sequences with superior properties. To address these challenges, future opportunities lie in pursuing advanced machine learning architectures, integrating multimodal data to provide richer contexts, developing active learning approaches to guide experimental efforts, and exploring transfer learning and unsupervised learning techniques. Many CAPE participants have already utilized some of these strategies, and we anticipate that even greater innovation will emerge in future competitions.

CONCLUSIONS

By moving computing and experimentation to the cloud, the CAPE challenge represents a unique, openly accessible platform for participants, regardless of their background training and resource availability. By engaging student teams and fostering a competitive, yet collaborative environment, CAPE successfully harnessed collective intelligence to explore the vast space of protein design. The stepwise improvement across successive CAPE rounds highlights the value of iterative learning and the potential for continued advancement in future challenges. Looking ahead, CAPE plans to expand its scope to include different protein categories, more assays and conditions, and a variety of sequence libraries, better reflecting the versatility of protein functions and applications.[9] For example, the third CAPE challenge held in 2024 (Table S1) focused on the mutant design of green fluorescence protein, and the results will be described in a later publication. Future collaboration with industry partners could help students engage with industry mentors and gain insight into real-world challenges. In the long run, CAPE aims to generate a diverse array of open data sets derived from standardized robotic assays, which can serve as well-structured, high-quality benchmarks to appraise ever-improving algorithms and models in this field. We envision CAPE, together with other community-driven efforts,[17,18] can accelerate progress toward the "AlphaFold moment" for protein design and engineering.[19] The ultimate success of CAPE will be measured by its ability to foster superior sequence−function predictors, innovative design algorithms, and, most importantly, next-generation protein engineers.

MATERIALS AND METHODS

Experimental Methods. All molecular biology, microbiology, and analytical chemistry protocols were performed by using previously reported robotic protocols on a biofoundry work cell.[15] Briefly, the Golden Gate Assembly was applied for rapid construction of combinatorial mutagenesis sequences from pre-existing gene fragments. The *rhlA* gene was divided into six segments: 1−227 bp, 224−381 bp, 378−435 bp, 432−494 bp, 491−524 bp, and 421−888 bp, which encoded protein fragments containing residues Arg74, Ala101, Gln143, Leu148, Ser173, and Gln176, respectively. The split sites and linker sequences were designed using the NEB Golden Gate Assembly Tool (https://goldengate.neb.com/). Each *rhlA* gene fragment was cloned into a receiver vector, Lv0-*ccdb*, using Golden Gate Assembly with *Bsm*BI. The resultant level-0 plasmids were used as the wild-type (WT) templates for site-directed mutagenesis, which was introduced by Gibson Assembly of two overlapping PCR products. In total, 120 level-0 substrate plasmids were constructed for the six target residues and stored in a single 384-well plate. For a specific target sequence, six corresponding substrate plasmids were cherry-picked using an Echo acoustic lipid handler and combined with the receiver plasmid pRSF-Duet-*rhlB*-ccdb for Golden Gate Assembly with *Bsa*I in 96-well PCR plates. The resulting reaction mixtures were used to transform chemically competent *E. coli* DH5*α* by heat shock. Generally, screening two random transformants was sufficient to identify at least one correct construct via Sanger sequencing.

To assess the inducible production and detection of monorhamnolipids, the mutant plasmids were transformed into *E. coli* BL21(DE3) cells, and three random colonies were transferred into 2 mL of LB medium in 24-deep well plates. After 16 h of cultivation at 37 °C, 20 *μ*L of stationary-phase cultures was used to inoculate 2 mL of fresh LB-Kan medium and incubated at 37 °C for 3−4 h to reach the exponential growth phase (OD_{600} = 0.6−0.8). IPTG was then added to a final concentration of 1 mM for inducible mono-RL production at 30 °C for 24 h.

For mass spectrometric (MS) analysis, the resulting cultures were mixed with an equal volume of methanol solution, followed by centrifugation at 4000 rpm for 10 min. The top phase was filtered through a 0.22 *μ*m filter membrane and collected into a 96-well plate. A C18 cartridge (G9205A, Agilent) was used for the RapidFire high-throughput MS system (Agilent). Eluent A (acetonitrile:H_2O (1:9, v/v)) and eluent B (acetonitrile:H_2O (9:1, v/v)) mixed with 5 mM ammonium acetate were used as mobile phases. The mass spectrometer was equipped with an electrospray ionization (ESI) source and operated in negative mode with a capillary voltage of 5 kV and a cone voltage of 30 V. Nitrogen was used as the nebulizer gas, and the source temperature was maintained at 120 °C. Quantification was performed using the multiple reaction monitoring (MRM) mode, with transitions of *m*/*z* 475 → 169 (Rha-(C_8−C_{10})) and *m*/*z* 503 → 169 (Rha-(C_{10}−C_{10})) for monorhamnolipid molecules.

Computing Method. For the Kaggle leaderboard, the Spearman correlation coefficient was used to rank competing teams based on their submitted prediction (*X*) on the Round 1 data set (*Y*) as a hidden test. The formula is as follows:

$$\rho = \frac{n\sum R(X_i)R(Y_i) - \sum R(X_i)\sum R(Y_i)}{\sqrt{n\sum R(X_i)^2 - (\sum R(X_i))^2}\cdot\sqrt{n\sum R(Y_i)^2 - (2\sum R(Y_i))^2}} \tag{1}$$

where *ρ* is the Spearman correlation coefficient, *n* is the number of sequences, and $R(X_i)$ and $R(Y_i)$ are the ranking in *X* and *Y* for sequence *i*, respectively. A baseline algorithm was included in the leaderboard using one-hot encoding for sequence representation and the Random Forest algorithm for sequence−function modeling, achieving a Spearman's *ρ* of 0.736.

The Shannon diversity index was used as a measure of amino acid diversity across sequence sets. The proportion of each amino acid of the total is calculated, and the Shannon entropy is obtained by taking the negative of the sum of each proportion multiplied by the binary log of that proportion. The Shannon diversity index is the average Shannon entropy of each site. The formula is as follows:

$$H' = \frac{1}{m}\sum_{j=1}^{m}\left(-\sum_{i=1}^{n} p_i \log_2 p_i\right) \tag{2}$$

where *H*′ is the Shannon diversity index, *m* (=6) is the number of amino acid sites, *n* (=20) is the number of distinct amino acids, and p_i is the proportion of each amino acid at a particular sequence position.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acssynbio.4c00588.

Figure S1, Experimental function evaluation of RhlA mutants designed by CAPE participants; Figure S2, winning strategies and model design in the CAPE challenge; Table S1, the key dates and events of the CAPE challenge; Table S2, teams, models, and scores in the inaugural CAPE challenge; and Table S3, teams, models, and scores in the second CAPE challenge (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Tong Si − *CAS Key Laboratory for Quantitative Engineering Biology, Shenzhen Institute of Synthetic Biology, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China; University of Chinese Academy of Sciences, Beijing 100049, China;* orcid.org/0000-0003-2985-9014; Email: tong.si@siat.ac.cn

Xiao Yi − *CAS Key Laboratory for Quantitative Engineering Biology, Shenzhen Institute of Synthetic Biology, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China; University of Chinese Academy of Sciences, Beijing 100049, China;* orcid.org/0000-0003-4025-856X; Email: xiao.yi@siat.ac.cn

Authors

Lihao Fu − *CAS Key Laboratory for Quantitative Engineering Biology, Shenzhen Institute of Synthetic Biology, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China*

Yuan Gao − *CAS Key Laboratory for Quantitative Engineering Biology, Shenzhen Institute of Synthetic Biology, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China; University of Chinese Academy of Sciences, Beijing 100049, China*

Yongcan Chen − *CAS Key Laboratory for Quantitative Engineering Biology, Shenzhen Institute of Synthetic Biology, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China*

Yanjing Wang − *CAS Key Laboratory for Quantitative Engineering Biology, Shenzhen Institute of Synthetic Biology, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China*

Xiaoting Fang − *CAS Key Laboratory for Quantitative Engineering Biology, Shenzhen Institute of Synthetic Biology, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China*

Shujun Tian − *CAS Key Laboratory for Quantitative Engineering Biology, Shenzhen Institute of Synthetic Biology, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China*

Hao Dong − *Kuang Yaming Honors School, Nanjing University, Nanjing 210023, China; State Key Laboratory of Analytical Chemistry for Life Science, Chemistry and Biomedicine Innovation Center (ChemBIC), Institute for Brain Sciences, Nanjing University, Nanjing 210023, China;* orcid.org/0000-0001-7280-7506

Yijian Zhang − *Kuang Yaming Honors School, Nanjing University, Nanjing 210023, China*

Zichuan Chen − *Kuang Yaming Honors School, Nanjing University, Nanjing 210023, China;* orcid.org/0009-0008-9623-3840

Zechen Wang − *School of Physics, Shandong University, Jinan 250100, China; Shanghai Zelixir Biotech Co. Ltd., Shanghai 200100, China;* orcid.org/0000-0003-4554-133X

Shantong Hu − *College of Life Science and Technology, Beijing University of Chemical Technology, Beijing 100049, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acssynbio.4c00588

Author Contributions

T.S. and X.Y. conceived and designed the CAPE challenge. X.Y. supervised organizational activities with help from Y.G. L.F. and Y.W. performed experimental validation. Y.G. and Y.C. examined computational models and developed the Kaggle competition. L.F., Y.C., and Y.G. analyzed experimental data. H.D., Y.Z., Z.C., Z.W., and S.H. led the winning teams. All authors participated in manuscript writing.

Author Contributions

●L.F., Y.G., and Y.C. contributed equally.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work is supported by the National Key Research and Development Program of China (2021YFA0910800), the Shenzhen Institute for Synthetic Biology, and the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB0480000). We acknowledge the Shenzhen Infrastructure for Synthetic Biology for instrument support and technical assistance. We thank all CAPE participants for their contributions and agreement to the open-access deposition of models and data on Kaggle and GitHub.

■ REFERENCES

(1) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Zídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583−589.

(2) Wang, C. Y.; Chang, P. M.; Ary, M. L.; Allen, B. D.; Chica, R. A.; Mayo, S. L.; Olafson, B. D. ProtaBank: A repository for protein design and engineering data. *Protein Sci.* **2018**, *27* (6), 1113−1124.

(3) Dallago, C. M. J.; Johnston, K. E.; Wittmann, B. J.; Bhattacharya, N.; Goldman, S.; Madani, A.; Yang, K. K. FLIP: Benchmark tasks in fitness landscape inference for proteins. *bioRxiv* **2021**.

(4) Moult, J.; Pedersen, J. T.; Judson, R.; Fidelis, K. A large-scale experiment to assess protein structure prediction methods. *Proteins* **1995**, *23* (3), ii−iv.

(5) Armer, C.; Letronne, F.; DeBenedictis, E. Support academic access to automated cloud labs to improve reproducibility. *Plos Biol.* **2023**, *21* (1), No. e3001919.

(6) Chao, R.; Mishra, S.; Si, T.; Zhao, H. Engineering biological systems using automated biofoundries. *Metab. Eng.* **2017**, *42*, 98−108.

(7) Hossain, Z.; Bumbacher, E. W.; Chung, A. M.; Kim, H.; Litton, C.; Walter, A. D.; Pradhan, S. N.; Jona, K.; Blikstein, P.; Riedel-Kruse, I. H. Interactive and scalable biology cloud experimentation for scientific inquiry and education. *Nat. Biotechnol.* **2016**, *34* (12), 1293−1298.

(8) Cooper, S.; Khatib, F.; Treuille, A.; Barbero, J.; Lee, J.; Beenen, M.; Leaver-Fay, A.; Baker, D.; Popovic, Z.; Players, F. Predicting protein structures with a multiplayer online game. *Nature* **2010**, *466* (7307), 756−60.

(9) Gomez-Hinostroza, E. S.; Gurdo, N.; Alvan Vargas, M. V. G.; Nikel, P. I.; Guazzaroni, M. E.; Guaman, L. P.; Castillo Cornejo, D. J.; Platero, R.; Barba-Ostria, C. Current landscape and future directions of synthetic biology in South America. *Front Bioeng Biotechnol* **2023**, *11*, 1069628.

(10) Friedberg, I.; Wass, M. N.; Mooney, S. D.; Radivojac, P. Ten simple rules for a community computational challenge. *Plos Comput. Biol.* **2015**, *11* (4), No. e1004150.

(11) Das, R.; Keep, B.; Washington, P.; Riedel-Kruse, I. H. Scientific discovery games for biomedical research. *Annu. Rev. Biomed Da S* **2019**, *2*, 253−279.

(12) Beal, J.; Telmer, C. A.; Vignoni, A.; Boada, Y.; Baldwin, G. S.; Hallett, L.; Lee, T.; Selvarajah, V.; Billerbeck, S.; Brown, B.; Cai, G. N.; Cai, L.; Eisenstein, E.; Kiga, D.; Ross, D.; Alperovich, N.; Sprent, N.; Thompson, J.; Young, E. M.; Endy, D.; Haddock-Angelli, T. Multicolor plate reader fluorescence calibration. *Synth Biol. (Oxf)* **2022**, *7* (1), ysac010.

(13) Toyoda, T. Methods for open innovation on a genome-design platform associating scientific, commercial, and educational communities in synthetic biology. *Methods Enzymol* **2011**, *498*, 189−203.

(14) Cyranoski, D. Synthetic-biology competition launches. *Nature* **2010**, DOI: 10.1038/news.2010.271.

(15) Hu, R.; Fu, L.; Chen, Y.; Chen, J.; Qiao, Y.; Si, T. Protein engineering via Bayesian optimization-guided evolutionary algorithm and robotic experiments. *Brief. Bioinform.* **2023**, *24* (1), bbac570.

(16) Protein mutation engineering competition phase2, 2023; https://kaggle.com/competitions/siatprotein2023.

(17) Armer, C. K. H.; Cortade, D. L.; Redestig, H.; Estell, D. A.; Yusuf, A.; Rollins, N.; Spinner, H.; Marks, D.; Brunette, T. J.; Kelly, P. J.; DeBenedictis, E. Results of the protein engineering tournament: an open science benchmark for protein modeling and design. *bioRxiv* **2024**.

(18) Koepnick, B.; Flatten, J.; Husain, T.; Ford, A.; Silva, D. A.; Bick, M. J.; Bauer, A.; Liu, G.; Ishida, Y.; Boykov, A.; Estep, R. D.; Kleinfelter, S.; Norgard-Solano, T.; Wei, L.; Players, F.; Montelione, G. T.; DiMaio, F.; Popovic, Z.; Khatib, F.; Cooper, S.; Baker, D. De novo protein design by citizen scientists. *Nature* **2019**, *570* (7761), 390−394.

(19) Chica, R. A.; Ferruz, N. What does it take for an 'AlphaFold Moment' in functional protein engineering and design? *Nat. Biotechnol.* **2024**, *42* (2), 173−174.